Bad Communities with High Modularity

Ath. Kehagias
Faculty of Engineering
Aristotle Univ. of Thessaloniki
kehagiat@auth.gr

Abstract

It is well known that Newman's modularity function Q_N has the form $Q_N = Q_d - Q_0$, where Q_d is the *intracluster edge density* and Q_0 is a term corresponding to the *null model*. Hence modularity maximization is influenced by Q_d, which favors a small number of clusters, and $-Q_0$ which favors balanced clusters. We show that the $-Q_0$ term can cause not only underestimation of the cluster number (the well known "resolution limit" of modularity) but, in certain cases, also *over*estimation. Furthermore, we construct families of graphs, each of which has a "natural" community structure which, however, does not maximize modularity. In fact, we show that we can always find a graph G with a "natural clustering" $\mathbf{V}$ and a sequence of clusterings $\mathbf{U}_x$ (with approximately equal-sized clusters) such that the pair $(G, \mathbf{U}_x)$ has higher modularity than $(G, \mathbf{V})$. More specifically, the pair $(G, \mathbf{V})$ has low "natural modularity" while the pair $(\mathbf{U}_x, G)$, by appropriate choice of x, can achieve modularity arbitrarily close to one. In addition, $\mathbf{U}_x$ can be arbitrarily different from the natural clustering $\mathbf{V}$; more specifically, by appropriate choice of x, their Jaccard similarity can become arbitrarily close to zero.

1 Introduction

Newman's *modularity function* Q_N is probably the most popular quality function in the community detection literature. A large number of community detection algorithms are based on some form of modularity maximization. However, this approach can also run into problems, as has been widely reported in the literature. This paper studies one particular way in which modularity can fail, and can be summarized as follows.

We will use the term "*cluster*" as a synonym of "community"; a *clustering* is a partition of the nodes of a graph; additional nomenclature and notation are introduced in Section 2.

Section 3 is devoted to an interpretation of Q_N. Modularity can be written in the form $Q_N = Q_d - Q_0$, where Q_d is the *intracluster edge density* and Q_0 is a term corresponding to the *null model*. It follows that maximizing modularity is equivalent to maximizing the sum of Q_d and $-Q_0$. As explained in Section 3.2, the term Q_d favors clusterings with few clusters and many intracluster edges; as explained in Section 3.3, the term $-Q_0$ favors "balanced clusterings" (i.e., clusterings with equally sized clusters) and is also responsible for *cluster number selection*. The fact that modularity maximization yields an estimate of the number of clusters is generally

perceived as an advantage over other quality functions. However, it is well known that this estimate can, in certain cases, be wrong. The literature has concentrated on the case where the estimate is lower than the true number of clusters ("*resolution limit*" of modularity) but we give examples of the opposite effect. These problems can be attributed to the previously mentioned fact that $-Q_0$ favors balanced clusterings.

In Section 4 we exploit the behavior of $-Q_0$ and construct examples in which modularity maximization yields arbitrarily inaccurate clusterings. More specifically, we construct families of graphs $\{G_{K,N_1,N_2}\}_{K=1}^{\infty}$ (where N_1, N_2 are parameters of the graph) with the following properties.

1. Each graph G_{K,N_1,N_2} has a "natural" clustering $\mathbf{V}_{K,N_1,N_2}$ which, however, does not maximize modularity.
2. We can find graphs G_{K,N_1,N_2} and clustering $\mathbf{U}_{K,,N_1,N_2,J}$ such that:
   2.1. the pair $(G_{K,N_1,N_2}, \mathbf{U}_{K,,N_1,N_2,J})$ has higher modularity than $(G_{K,N_1,N_2}, \mathbf{V}_{K,,N_1,N_2})$;
   2.2. the modularity of $(G_{K,N_1,N_2}, \mathbf{U}_{K,,N_1,N_2,J})$ is arbitrarily close to one;
   2.3. the similarity between clusterings $\mathbf{V}_{K,N_1,N_2}$ and $\mathbf{U}_{K,N_1,N_2,J}$ is arbitrarily close to zero.

In Section 5 we discuss our results and (previously published) related work.

2 Preliminaries

1. We deal with finite graphs without multiple edges. A *graph* G is a pair (V, E), where V is the *node set* (we will always assume $V = \{1, 2, ..., n\}$; hence the number of nodes is $n = |V|$) and $E \subseteq V \times V$ is the *edge set* (and $m = |E|$ is the number of edges).
2. The *adjacency matrix* of G is an $n \times n$ matrix A with $A_{m,n} = 1$ iff $\{m, n\} \in E$ and 0 otherwise. There is a one-to-one correspondence between a graph G and its adjacency matrix A.
3. A *clustering* of $G = (V, E)$ is a partition $\mathbf{V}=\{V_1, ..., V_K\}$ of V (i.e., $\cup_{k=1}^{K} V_k = V$ and $\forall k, l : V_k \cap V_l = \emptyset$). The *size* of the clustering is K, the number of clusters. Given a graph $G = (V, E)$, we denote by $\mathcal{V}$ the set of all clusterings of V and by $\mathcal{V}_K$ the set of clusterings of size K.
4. Each of the V_k's is called a *cluster* or, synonymously, a *community*. For $k = 1, 2, ..., K$ we set $n_k = |V_k|$ and $n'_k = |V - V_k|$. We have $\sum_{k=1}^{K} n_k = n$.
5. A clustering $\mathbf{V}=\{V_1, ..., V_K\}$ of the graph $G = (V, E)$, induces a family of edge sets $\mathbf{E}=\{E_{11}, E_{12}, ..., E_{KK}\}$, where

   $$e = \{u, v\} \in E_{ij} \text{ iff } u \in V_i \text{ and } v \in V_j.$$

   Note that $E_{ij} = E_{ji}$. The elements of E_{ii} (for $i \in \{1, ..., n\}$) will be called *intracluster* edges; the elements of E_{ij} (for $i \neq j$) will be called *extracluster* edges. We also write $E_k = E_{kk}$.

6. The *degree function* is denoted by $\deg(\cdot)$ and is defined as follows: for any $v \in V$, $\deg(v) = |\{e : v \in e\}|$ is the number of edges incident on v; for any $U \subseteq V$, $\deg(U) = \sum_{v \in U} \deg(v)$, the sum of degrees of the nodes contained in U.

7. The *Jaccard similarity* index is defined as follows. Given any two clusterings $\mathbf{W}_1$, $\mathbf{W}_2$ define

   a_{11} = "num. of node pairs (i,j) in same cluster under $\mathbf{W}_1$ and same cluster under $\mathbf{W}_2$";
   a_{10} = "num. of node pairs (i,j) in same cluster under $\mathbf{W}_1$ and different cluster under $\mathbf{W}_2$";
   a_{01} = "num. of node pairs (i,j) in different cluster under $\mathbf{W}_1$ and same cluster under $\mathbf{W}_2$";
   a_{00} = "num. of node pairs (i,j) in different cluster under $\mathbf{W}_1$ and different cluster under $\mathbf{W}_2$".

   Then the *Jaccard similarity* index $S(\mathbf{W}_1, \mathbf{W}_2|G)$ (with respect to the graph G) is

   $$S(\mathbf{W}_1, \mathbf{W}_2|G) = \frac{a_{11}}{a_{10} + a_{01} + a_{11}}.$$

   $S(\mathbf{W}_1, \mathbf{W}_2|G)$ takes values in $[0,1]$; values close to 1 show that $\mathbf{W}_1, \mathbf{W}_2$ are very similar; values close to 0 that they are very different. Note that the similarity of $\mathbf{W}_1$ and $\mathbf{W}_2$ is computed with respect to G.

3 An Intepretation of Modularity

3.1 Modularity

Given a graph $G = (V, E)$ with adjacency matrix A, we denote the modularity of a clustering $\mathbf{V}$ by $Q_N(\mathbf{V}, G)$ and, following [25], we define it by

$$Q_N(\mathbf{V}, G) = \frac{1}{2m} \sum_{i,j \in V} \left(A_{ij} - \frac{\deg(i)\deg(j)}{2m} \right) \delta(i,j). \tag{1}$$

Our notation emphasizes that $Q_N(\mathbf{V}, G)$ is a function of both the graph and the clustering.

The motivation for introducing modularity can be seen by the following interpretation[1]: $Q_N(\mathbf{V}, G)$ *is the difference of the fraction of the intracluster edges minus the expected value of the same quantity in a graph* (the null model) *with the same clusters but random connections between the nodes.* A large value of $Q_N(\mathbf{V}, G)$ indicates that, under $\mathbf{V}$, G is quite different from the null model; this is taken as evidence of G having "*strong community structure*" which is "well captured" by $\mathbf{V}$. Hence modularity is a *clustering quality function.*

Other interpretations of modularity are possible; we will propose one a little later. But first let us note that, in addition to characterizing a single $(\mathbf{V}, G)$ pair, modularity can be used to *compare* clusterings: *by definition*, $\mathbf{V}$ is a better clustering of G than $\mathbf{V}'$ iff $Q_N(\mathbf{V}, G) > Q_N(\mathbf{V}', G)$. Taking this one step further, $\mathbf{V}^* = \arg\max_{\mathbf{V}} Q_N(\mathbf{V}, G)$ is the *best* clustering of G, hence modularity maximization can be used to obtain graph clusterings (i.e., perform community detection); and a large value of $\max_{\mathbf{V}} Q_N(\mathbf{V}, G)$ indicates that G has strong community structure.

[1] Which is a slight paraphrase of Newman and Girvan [25].

While modularity is widely used, its shortcomings have been widely reported in the literature. For example, the *modularity resolution limit* has attracted a lot of attention [13, 14]; we will discuss it in Section 3.3. But first let us note what appears to be a more basic limitation of modularity. As already mentioned, a large $Q_N(\mathbf{V}, G)$ value indicates strong community structure and good clustering; but what *is* a "large $Q_N(\mathbf{V}, G)$ value"? While it is known [7] that $-\frac{1}{2} \leq Q_N(\mathbf{V}, G) \leq 1$, many examples appear in the community detection literature [13, 12, 14] which have strong (intuitively perceived) community structure and yet their maximum modularity is closer to zero than to one. The converse can also be true[12, 4] [2].

A frequently proposed explanation for the shortcomings of modularity is that the null model assumption is not justified[3] [12]. In Section 3.3 we will consider an alternative explanation. But first we will examine another clustering quality function.

3.2 Intracluster Edge Density

A popular characterization of a network / graph community is that "*there must be more edges 'inside' the community than edges linking vertices of the community with the rest of the graph*" [12, Section B.1]. We will take this as a basic guiding principle[4].

A *prima facie* reasonable way to quantify the principle is through the *intracluster edge density*, denoted by $Q_d(\mathbf{V}, G)$ and defined by

$$Q_d(\mathbf{V}, G) = \frac{\sum_{k=1}^{K} |E_k|}{m}. \tag{2}$$

For every G and $\mathbf{V}$, $Q_d(\mathbf{V}, G) \in [0, 1]$. A high (i.e., close to 1) value of $Q_d(\mathbf{V}, G)$ indicates that the pair $(\mathbf{V}, G)$ has many intracluster and few extracluster edges.

Unfortunately, a high $Q_d(\mathbf{V}, G)$ value does not guarantee either that G has strong community structure or that $\mathbf{V}$ is a good clustering of G. Indeed we can always achieve $Q_d(\mathbf{V}, G) = 1$ by taking $\mathbf{V} = \{V\}$ (i.e., the unique clustering of size one) but this tells us nothing about the "true" community structure of G. This observation can be generalized. First define the following function

$$F_G(K) = \max_{\mathbf{V} \in \mathcal{V}_K} Q_d(\mathbf{V}, G). \tag{3}$$

For a given graph G, $F_G(K)$ is the maximum intracluster edge density achieved by clusterings of size K. Now we can prove the following.

Theorem 3.1 *For any graph $G = (V, E)$, $F_G(K)$ is a nonincreasing function of K.*

Proof. There exists a single clustering of size one, namely $\mathbf{V}^{(1)} = \{V\}$. Denote the intracluster edges by $E^{(1)}$; obviously $E_1^{(1)} = E$ (i.e., *all* edges are intracluster). Hence $F_G(1) = \frac{\left|E_1^{(1)}\right|}{|E|} = 1$.

[2] In [12] is stated that ""In Section VI.C we have seen that high values of the modularity of Newman and Girvan do not necessarily indicate that a graph has a definite cluster structure". And in [4]: "since fluctuations can induce high modularity in random graphs, one must always approach the raw magnitude of Q with caution".

[3] In [12] is stated that "The weak point of the null model is the implicit assumption that each vertex can interact with every other vertex, which implies that each part of the graph knows about everything else."

[4] An extreme statement of this principle appears in [8]: "a community network $G_0 = (V; E_0)$ [is] a graph G_0 that is a disjoint union of complete subgraphs".

Let $\mathbf{V}^{(2)}=\left\{V_1^{(2)}, V_2^{(2)}\right\}$ be the optimal clustering of size two; the intracluster edges are $E_1^{(2)}$ and $E_2^{(2)}$. We have $E_1^{(2)} \cup E_2^{(2)} \subseteq E$ and $\left|E_1^{(2)}\right| + \left|E_2^{(2)}\right| \leq |E|$. Hence

$$F_G(2) = Q_d\left(\mathbf{V}^{(2)}, G\right) = \frac{\sum_{k=1}^{2}\left|E_k^{(2)}\right|}{|E|} \leq 1 = F_G(1).$$

Let $\mathbf{V}^{(3)} = \left\{V_1^{(3)}, V_2^{(3)}, V_3^{(3)}\right\}$ be the optimal clustering of size three; the intracluster edges willl be $E_1^{(3)}$, $E_2^{(3)}$, $E_3^{(3)}$. Create the clustering (of size two) $\mathbf{V}'=\left\{V_1^{(3)}, V_2^{(3)} \cup V_3^{(3)}\right\}$ with intracluster edges E_1', E_2'. Note that $E_1^{(3)} = E_1'$ and $E_2^{(3)} \cup E_3^{(3)} \subseteq E_2'$. In other words, some edges which were extracluster in $\mathbf{V}^{(3)}$ may become intracluster in $\mathbf{V}'$. Hence we have

$$\left|E_1^{(3)}\right| = |E_1'|, \qquad \left|E_2^{(3)}\right| + \left|E_3^{(3)}\right| \leq |E_2'|$$

and

$$F_G(3) = Q_d\left(\mathbf{V}^{(3)}, G\right) = \frac{\sum_{k=1}^{3}\left|E_k^{(3)}\right|}{|E|} \leq \frac{\sum_{k=1}^{2}|E_k'|}{|E|} = Q_d\left(\mathbf{V}', G\right) \leq Q_d\left(\mathbf{V}^{(2)}, G\right) \leq F_G(2).$$

Proceeding in this manner we get $1 = F_G(1) \geq F_G(2) \geq ... \geq F_G(n) \geq 0$ and the proof is complete. ■

Hence for any G the optimal K (with respect to Q_d) is $K = 1$; it follows that Q_d maximization cannot determine the optimal number of clusters. However, if K is given in advance, then $\mathbf{V}^{(K)} = \arg\max_{\mathbf{V}\in\mathcal{V}_K} Q_d(\mathbf{V}, G)$ is a reasonable candidate for the best clustering of size K. Actually this has often been expressed as a criticism of intracluster edge density, e.g., in [12] it is stated that "*Algorithms for graph partitioning are not good for community detection, because it is necessary to provide as input the number of groups...*". However, this criticism is valid only to the extent that other algorithms exist which *can* obtain the the number of groups (clusters).

3.3 Modularity as Augmented Intracluster Edge Density

As seen in Section 3.1, an alleged advantage of modularity is that its maximization yields the correct number of clusters; however, it is well known that there is a modularity *resolution limit* [13, 14], i.e., in certain cases the maximum modularity clustering yields the wrong cluster number. Why is this the case?

Our explanation will make use of the modularity formula

$$Q_N(\mathbf{V}, G) = \sum_{k=1}^{K}\frac{|E_k|}{m} - \sum_{k=1}^{K}\left(\frac{\deg(V_k)}{2m}\right)^2 \tag{4}$$

which, as is well known, is equivalent to (1). Comparing (4) and (2) we see that

$$Q_N(\mathbf{V}, G) = Q_d(\mathbf{V}, G) - Q_0(\mathbf{V}, G) \tag{5}$$

where

$$Q_0(\mathbf{V}, G) = \sum_{k=1}^{K} \left(\frac{\deg(V_k)}{2m} \right)^2. \tag{6}$$

Hence modularity can be seen as an *augmented intracluster edge density* or, in other words, the difference of intracluster edge density $Q_d(\mathbf{V}, G)$ and the auxiliary function $Q_0(\mathbf{V}, G)$. What is the role of $Q_0(\mathbf{V}, G)$? Its introduction is usually justified with reference to the null model [25]. Here is another way to look at the matter.

Suppose momentarily that K is given and we want to minimize $Q_0(\mathbf{V}, G) = \sum_{k=1}^{K} \left(\frac{\deg(V_k)}{2m} \right)^2$. For simplicity of notation define $p_k = \frac{\deg(V_k)}{2m}$; we have

$$\sum_{k=1}^{K} p_k = \sum_{k=1}^{K} \frac{\deg(V_k)}{2m} = 1.$$

Hence we want to solve the following problem:

$$\text{given } K, \text{ minimize } \sum_{k=1}^{K} p_k^2 \text{ subject to: } 0 \leq p_k \leq 1 \text{ and } \sum_{k=1}^{K} p_k = 1. \tag{7}$$

Assuming for the moment that the p_k's are continuously valued (this will be later relaxed), the solution to (7) is $p_k = \frac{1}{K}$ for all k; the minimum thus achieved is $\frac{1}{K}$. If K is not given, the problem becomes

$$\text{minimize } \sum_{k=1}^{K} p_k^2 \text{ subject to: } K \in \{1, ..., n\}, 0 \leq p_k \leq 1 \text{ and } \sum_{k=1}^{K} p_k = 1. \tag{8}$$

Solving (8) separately for each $K \in \{1, ..., n\}$ we see that the overall optimal solution is $K = n$ (each cluster contains a single node) and $p_k = \frac{1}{n}$ for all k; the minimum thus achieved is $\frac{1}{n}$.

Let us now return to modularity maximization. We know that (a) $Q_N = Q_d - Q_0$, (b) Q_d achieves its maximum at $K = 1$ and (c) Q_0 achieves its minimum at $K = n$. At an intuitive level, it is easy to understand the factors influencing the outcome of modularity maximization: the Q_d term pulls K towards small values and the $-Q_0$ towards large ones; in addition the Q_d term favors clusterings which correspond to the "natural" community structure of G, while the $-Q_0$ favors "balanced" clusterings in which all clusters contain more or less the same number of nodes. These conclusions are based on the maximization problem (8), in which the p_k's vary continuously. It is reasonable to expect that, for large n, they will also be (at least approximately) true when p_k takes discrete values of the form $p_k = \frac{\deg(V_k)}{2M}$. This can be justified as follows: with a large V, $\mathcal{V}$ is also large and it becomes likely that some clustering $\mathbf{V} = \{V_1, ..., V_K\}$ exists which can make the terms $\frac{\deg(V_k)}{2m}$ approximately equal for all k (of course, this will also depend on the distribution of the degrees $\deg(v)$, since $\deg(V_k) = \sum_{v \in V_k} \deg(v)$). Hence intuitively we expect that the role of the Q_0 term is to (a) increase the number of clusters and (b) equalize the sizes of the clusters.

Similar remarks have appeared in previous works. For example in [37] is stated that "the existing modularity optimization method does not perform well in the presence of unbalanced

community structures" and in [13] is stated that "For modularity's null model graphs, the modularity maximum corresponds to an equipartition of the graph". An interesting point is that the resolution limit has been identified as the tendency of modularity maximization to underestimate the "true" number of clusters[5]. But, as we will see in Section 4, modularity maximization may also produce the opposite effect, i.e., *over*estimation of cluster number.

One method used to address the resolution limit is to introduce a modified modularity function. This function can often [18, 22, 26, 33] be written in the form

$$Q_N^{\gamma}(\mathbf{V}, G) = Q_d(\mathbf{V}, G) - \gamma Q_0(\mathbf{V}, G)$$

where γ is a "tuning parameter". With $\gamma = 1$, $Q_N^1(\mathbf{V}, G) = Q_N(\mathbf{V}, G)$, the original Newman's modularity. If this underestimates (resp. overestimates) the "true" number of clusters, formation of more clusters can be encouraged by increasing (resp. decreasing) γ and hence the influence of the $-Q_0(\mathbf{V}, G)$ term on the maximization problem.

4 Bad Clusterings with High Modularity

In this section we prove the existence of graphs which have (i) a "*natural*" clustering and (ii) a family of "*arbitrarily bad*" clusterings such that the arbitrarily bad clusterings achieve higher modularity than the natural one. In addition we show that the arbitrarily bad clusterings can achieve modularity arbitrarily close to one and they can be "*arbitrarily different*" from the natural clustering. (In the sequel we will explain precisely what we mean by the terms "natural", "arbitrarily bad" and "arbitrarily different".) We will establish all of these results using two parametric graph families. Obviously, these results indicate that, at least in certain cases, modularity is not a good quality function.

4.1 First Example

Let us construct a family of graphs G_{K,N_1,N_2} (K, N_1, N_2 are parameters) such that the following two properties are satisfied.

P1 For every K, N_1, N_2, the graph G_{K,N_1,N_2} has an easily recognized "natural" clustering $\mathbf{V}_{K,N_1,N_2}$.

P2 We can select K, N_1, N_2 and a clustering $\mathbf{U}^*_{K,N_1,N_2}$ (different from $\mathbf{V}_{K,N_1,N_2}$) such that

$$Q_N\left(\mathbf{U}^*_{K,N_1,N_2}, G_{K,N_1,N_2}\right) > Q_N\left(\mathbf{V}_{K,N_1,N_2}, G_{K,N_1,N_2}\right). \tag{9}$$

Suppose K, N_1, N_2 are given. Let us first define the disconnected graph G_{N_1,N_2} to be the union of a path of N_1 nodes and a path of N_2 nodes; and then let the disconnected graph G_{K,N_1,N_2} be the union of K disconnected copies of G_{N_1,N_2}. The construction is illustrated in Figure 1.

[5]For example in [13] is stated that "The networks that we have examined are fairly small but the problem we have discovered can only get worse if we increase the network size, especially when small communities coexist with large ones and the module size distribution is broad, which seems to happen in many cases"; in [12] is stated that "modules identified through modularity optimization may actually be combinations of smaller modules".

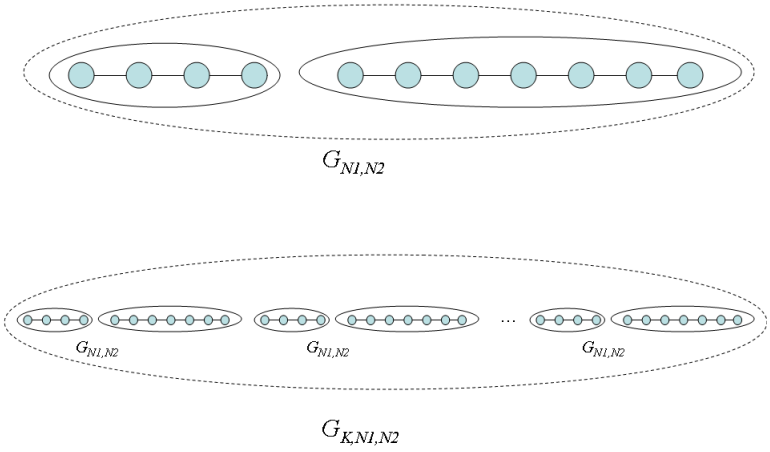

Figure 1: Graph Family A

What is the the *natural* clustering of G_{K,N_1,N_2}? We claim it is $\mathbf{V}_{K,N_1,N_2} = \{V_{K,N_1,N_2,1}, V_{K,N_1,N_2,2}, ..., V_{K,N_1,N_2,2K}\}$, where $V_{K,N_1,N_2,k}$ is the node set of the k-th disconnected component of G (with $k \in \{1, 2, ..., 2K\}$, see Figure 1). At the risk of belaboring the obvious, we note that, if $u \in V_{K,N_1,N_2,i}$ and $v \in V_{K,N_1,N_2,j}$ and $i \neq j$, then there exists no path connecting u and v; hence they should never be put in the same cluster. So the biggest possible clusters are the $V_{K,N_1,N_2,i}$'s. On the other hand, there is no justification for splitting any $V_{K,N_1,N_2,i}$ into smaller clusters, since all of its nodes have the same connectivity pattern. Hence $\mathbf{V}_{K,N_1,N_2}$ is the "intuitively best" (i.e., the "natural") clustering of G_{K,N_1,N_2}.

For every triple (K, N_1, N_2) let us now introduce a *sequence* $\{\mathbf{U}_{K,N_1,N_2,J}\}_{J=1}^{\infty}$ of clusterings of G_{K,N_1,N_2}. For a fixed J, let $L = \left\lfloor \frac{n}{J} \right\rfloor$; writing for brevity $\mathbf{U}_J$ in place of $\mathbf{U}_{K,N_1,N_2,J}$, we let $\mathbf{U}_J = \{U_1, ..., U_J, U_{J+1}\}$ consist of the following $J+1$ clusters:

$$U_1 = \{1, ..., L\}\,,\ U_2 = \{L+1, ..., 2L\}\,,\ ...\ ,\ U_J = \{(J-1)\,L+1, ..., JL\}\,,\ U_{J+1} = \{JL+1, ..., n\}\,;$$

if $n = JL$ then $U_{J+1} = \emptyset$. In other words, $\mathbf{U}_J$ contains J clusters each containing the same number, $L = \left\lfloor \frac{n}{J} \right\rfloor$, of nodes and perhaps an additional cluster with fewer than L nodes. Obviously $\mathbf{U}_J$ is a "well balanced" clustering.

Our goal is to prove (9). We will do this in three steps, corresponding to the following three propositions.

Lemma 4.1 *For every $K, N_1, N_2 \in \mathbb{N}$ with $N_1, N_2 \geq 3$ we have*

$$Q_N\left(\mathbf{V}_{K,N_1,N_2}, G_{K,N_1,N_2}\right) = 1 - \frac{(N_1-1)^2 + (N_2-1)^2}{K\,(N_1+N_2-2)^2}. \tag{10}$$

Proof. We fix K, N_1, N_2 and, for brevity, we write G for G_{K,N_1,N_2} and $\mathbf{V}$ for $\mathbf{V}_{K,N_1,N_2}$. We have

$$Q_N\left(\mathbf{V}, G\right) = \frac{\sum_{k=1}^{2K} |E_k|}{m} - \frac{\sum_{k=1}^{2K} \left(\deg\left(V_k\right)\right)^2}{(2m)^2}.$$

G has no extracluster edges under $\mathbf{V}$, hence we have

$$\frac{\sum_{k=1}^{2K}|E_k|}{m}=1. \tag{11}$$

We can separate $\mathbf{V}$ into two subsets of clusters: $\mathbf{V}'=\{V_1,V_3,...,V_{2K-1}\}$ contains the the clusters with N_1 nodes and $\mathbf{V}''=\{V_2,V_4,...,V_{2K}\}$ contains the the clusters with N_2 nodes. Each $V_k\in\mathbf{V}'$ has N_1-2 "inner nodes" of degree 2 and two "border nodes" of degree 1; similarly, each $V_k\in\mathbf{V}''$ has degree $\deg(V_k)=2(N_1-2)+2=2(N_2-1)$. Hence

$$\begin{aligned}&\forall: V_k\in\mathbf{V}':\deg(V_k)=2(N_1-2)+2=2(N_1-1)\\&\forall: V_k\in\mathbf{V}'':\deg(V_k)=2(N_2-2)+2=2(N_2-1)\end{aligned}$$

The total number of edges is

$$m=\frac{\sum_{V_k\in\mathbf{V}'}\deg(V_k)+\sum_{V_k\in\mathbf{V}''}\deg(V_k)}{2}=K(N_1+N_2-2).$$

Also,

$$\begin{aligned}\frac{\sum_{k=1}^{2K}(\deg(V_k))^2}{(2m^2)}&=\frac{\sum_{V_k\in\mathbf{V}'}(2(N_1-1))^2}{(2K(N_1+N_2-2))^2}+\frac{\sum_{V_k\in\mathbf{V}''}(2(N_2-1))^2}{(2K(N_1+N_2-2))^2}\\&=\frac{K(N_1-1)^2+K\cdot(N_2-1)^2}{K^2(N_1+N_2-2)^2}=\frac{(N_1-1)^2+(N_2-1)^2}{K(N_1+N_2-2)^2}.\end{aligned} \tag{12}$$

Combining (11) and (12) we get (10). ■

Lemma 4.2 *For every $K,N_1,N_2,J\in\mathbb{N}$ with $N_1,N_2\geq 3$ we have*

$$Q_N(\mathbf{U}_{K,N_1,N_2,J},G_{K,N_1,N_2})\geq 1-\frac{1}{K(N_1+N_2-2)}J-\frac{2(N_1+N_2)^2}{(N_1+N_2-2)^2}J^{-1}. \tag{13}$$

Proof. We write G for G_{K,N_1,N_2}, $\mathbf{V}$ for $\mathbf{V}_{K,N_1,N_2}$ and also $\mathbf{U}_J$ for $\mathbf{U}_{K,N_1,N_2,J}$. We have

$$Q_N(\mathbf{U}_J,G)=\frac{\sum_{k=1}^{J+1}|E_k|}{m}-\frac{\sum_{k=1}^{J+1}(\deg(U_k))^2}{(2m)^2}.$$

Consider first $\frac{\sum_{k=1}^{J+1}|E_k|}{m}$. A little thought shows that $\mathbf{U}_J$ has at most $J+1$ clusters and J extracluster edges. Hence

$$\forall J:\frac{\sum_{k=1}^{J+1}|E_k|}{m}\geq\frac{m-J}{m}=1-\frac{J}{m}=1-\frac{1}{K(N_1+N_2-2)}J. \tag{14}$$

Consider now $\frac{\sum_{k=1}^{J}(\deg(U_k))^2}{(2m)^2}$. Each U_k has no more than $\frac{n}{J}=\frac{K(N_1+N_2)}{J}$ nodes and each node has degree at most 2. Hence

$$\forall J:\frac{\sum_{k=1}^{J+1}(\deg(U_k))^2}{(2m)^2}\leq\frac{(J+1)\cdot\left(2\frac{K(N_1+N_2)}{J}\right)^2}{2^2K^2(N_1+N_2-2)^2}\leq\frac{2(N_1+N_2)^2}{(N_1+N_2-2)^2}J^{-1} \tag{15}$$

(since $\forall J \in \mathbb{N} : \frac{J+1}{J} \leq 2$). Combining (14) and (15) we get (13). ■

Hence, to ensure $Q_N\left(\mathbf{U}_{K,N_1N_2,J}, G_{K,N_1N_2}\right) > Q_N\left(\mathbf{V}_{K,N_1N_2}, G_{K,N_1N_2}\right)$ (i.e., that the natural clustering $\mathbf{V}_{K,N_1N_2}$ has lower modularity than $\mathbf{U}_{K,N_1N_2,J}$) it suffices to select K, N_1, N_2, J appropriately and use Lemmas 4.1 and 4.2. A sufficient condition, obtained from (10) and (13), is

$$1 - \frac{1}{K\left(N_1+N_2-2\right)}J - \frac{2\left(N_1+N_2\right)^2}{\left(N_1+N_2-2\right)^2}J^{-1} > 1 - \frac{\left(N_1-1\right)^2+\left(N_2-1\right)^2}{K\left(N_1+N_2-2\right)^2}. \tag{16}$$

Inspecting (16) we see that one way to satisfy it is by fixing N_1 and letting J be "sufficiently larger" than K and N_2 "sufficiently larger" than J. This is the main idea used in the proof of the following theorem.

Theorem 4.3 *For every $K \in \mathbb{N}$ and $\varepsilon \in \left(0, \frac{1}{K}\right)$ there exist $N_1, N_2, J \in \mathbb{N}$ (depending on ε and K) such that*

$$Q_N\left(\mathbf{U}_{K,N_1,N_2,J}, G_{K,N_1,N_2}\right) > 1-\varepsilon > 1 - \frac{1}{K} = Q_N\left(\mathbf{V}_{K,N_1,N_2}, G_{K,N_1,N_2}\right), \tag{17}$$

$$\varepsilon > S\left(\mathbf{U}_{K,N_1,N_2,J}, \mathbf{V}_{K,N_1,N_2} | G_{K,N_1,N_2}\right) \tag{18}$$

Proof. Take any K and $\varepsilon > \frac{1}{K}$ and let $N_1 = 3$, $J = xK$, $N_2 = x^2K$. To prove (17) note that

$$Q_N\left(\mathbf{U}_{K,N_1,N_2,J}, G_{K,N_1,N_2}\right) > 1 - \frac{x}{\left(1+x^2K\right)} - \frac{2\left(3+x^2K\right)^2}{\left(1+x^2K\right)^2 xK},$$

$$Q_N\left(\mathbf{V}_{K,N_1,N_2}, G_{K,N_1,N_2}\right) = 1 - \frac{4+\left(x^2K-1\right)^2}{K\left(1+x^2K\right)^2}.$$

We have

$$1 - \frac{xK}{\left(1+x^2K\right)} - \frac{2\left(3+x^2K\right)^2}{\left(1+x^2K\right)^2 xK} = 1 - \frac{3}{xK} + o\left(x\right),$$

$$1 - \frac{4+\left(x^2K-1\right)^2}{K\left(1+x^2K\right)^2} = 1 - \frac{1}{K} + o\left(x\right).$$

Hence, for (17) to hold, x must be big enough for $o\left(x\right)$ to be negligible and we also need

$$1 - \frac{3}{xK} > 1 - \varepsilon > 1 - \frac{1}{K}$$

which is satisfied for any $x > 3/K\varepsilon$. In short, we can satisfy (17) for every $K \in \mathbb{N}$ and and $\varepsilon \in \left(0, \frac{1}{K}\right)$, by taking x sufficiently big and $N_1 = 3$, $J = xK$, $N_2 = x^2K$.

Let us prove (18). Let b (resp. c) be the number of node pairs in the same cluster under $\mathbf{U}_{K,N_1,N_2,J}$ (resp. under $\mathbf{V}_{K,N_1,N_2}$). We obviously have $b = a_{01} + a_{11} \geq a_{11}$ and $a_{10} + a_{01} + a_{11} > a_{10} + a_{11} = c > 0$. Hence

$$S\left(\mathbf{U}_{K,N_1,N_2,J}, \mathbf{V}_{K,N_1,N_2} | G_{K,N_1,N_2}\right) = \frac{a_{11}}{a_{10}+a_{01}+a_{11}} \leq \frac{b}{c}.$$

We will obtain an upper bound for b. Since each U_j contains no more than $L = \frac{n}{J}$ nodes , the number of node pairs that can be formed in U_j is no more than $\frac{\left(\frac{n}{J}\right)\left(\frac{n}{J}-1\right)}{2} < \frac{n^2/2}{J^2}$. Also, $n = K(N_1 + N_2)$ so, for big N_2, $\frac{n^2/2}{J^2} < \frac{KN_2^2}{J^2}$. There are at most $J+1$ clusters, so we have

$$b \leq (J+1)\frac{(KN_2)^2}{J^2} = (xK+1)\frac{(Kx^2K)^2}{(xK)^2} = K^3x^3 + K^2x^2.$$

Now we will compute c. In $\mathbf{V}_{K,N_1,N_2}$ there exist K clusters of $N_1 = 4$ nodes; each has $\frac{N_1(N_1-1)}{2} = 6$ node pairs; there also exist K clusters of N_2 nodes; each has $\frac{N_2(N_2-1)}{2}$ node pairs. We have

$$c = 6K + K\frac{N_2(N_2-1)}{2} = 6K + K\frac{x^2K(x^2K-1)}{2} = \frac{1}{2}K^3x^4 - \frac{1}{2}K^2x^2 + 6K.$$

And so we have

$$0 \leq S(\mathbf{U}_{K,N_1,N_2,J}, \mathbf{V}_{K,N_1,N_2}|G_{K,N_1,N_2}) < \frac{K^3x^3 + K^2x^2}{\frac{1}{2}K^3x^4 - \frac{1}{2}K^2x^2 + 6K} \Rightarrow$$
$$0 \leq \lim_{x\to\infty} S(\mathbf{U}_{K,N_1,N_2,J}, \mathbf{V}_{K,N_1,N_2}|G_{K,N_1,N_2}) \leq \lim_{x\to\infty} \frac{K^3x^3 + K^2x^2}{\frac{1}{2}K^3x^4 - \frac{1}{2}K^2x^2 + 6K} = 0.$$

Hence, for x large enough (18) is satisfied. ■

Remark. We see from (17) that we can always find a clustering $\mathbf{U}_{K,N_1,N_2,J}$ which achieves modularity arbitrarily close to one and, in addition, better than the natural clustering $\mathbf{V}_{K,N_1,N_2}$. Note that here "close to one" means greater than $1-\varepsilon$, where ε can get arbitrarily small independently of K. On the other hand, $Q_N(\mathbf{V}_{K,N_1,N_2}, G_{K,N_1,N_2}) < 1 - \frac{1}{K}$ which can become rather small when K takes small values. In other words, G_{K,N_1,N_2} has low "natural modularity" (the one achieved by the pair $(\mathbf{V}_{K,N_1,N_2}, G_{K,N_1,N_2})$) but it can be assigned high "artificial modularity" (the one achieved by the pair $(\mathbf{U}_{K,N_1,N_2,J}, G_{K,N_1,N_2})$).

Remark. Furthermore, from (18) we see that $\mathbf{U}_{K,N_1,N_2,J}$ is very different from $\mathbf{V}_{K,N_1,N_2}$ (according to the Jaccard similarity; similar results can also be proved for other clustering similarity indices, for example the information theoretic distance used by Danon [9]). In a sense, this is not surprising: recall that $J = xK$ and we can choose x arbitrarily large; hence $\mathbf{U}_{K,N_1,N_2,J}$ will have xK clusters, which is many more than the $2K$ clusters contained in $\mathbf{V}_{K,N_1,N_2}$ (note that for the $\{G_{K,N_1,N_2}\}_{K=1}^{\infty}$ family modularity maximization *over*estimates the number of clusters).

Example. Let us give some numerical examples of the above ideas.

1. Taking $K = 2$, $x = 4$ and $N_1 = 3$ we also get $J = xK = 8$ and $N_2 = x^2K = 32$. With these values we get

$$Q_N(\mathbf{V}_{K,N_1,N_2}, G_{K,N_1,N_2}) = 0.5569 = 1 - \frac{(N_1-1)^2 + (N_2-1)^2}{K(N_1+N_2-2)^2},$$
$$Q_N(\mathbf{U}_{K,N_1,N_2,J}, G_{K,N_1,N_2}) = 0.7657 > 0.5976 = 1 - \frac{1}{K(N_1+N_2-2)}J - \frac{2(N_1+N_2)^2}{(N_1+N_2-2)^2}J^{-1};$$

both (16) and (17) are satisfied.

2. Taking $K = 4$, $x = 10$ and $N_1 = 3$ we also get $J = xK = 40$ and $N_2 = x^2K = 400$. With these values we get

$$Q_N\left(\mathbf{V}_{K,N_1,N_2}, G_{K,N_1,N_2}\right) = 0.7525 = 1 - \frac{(N_1-1)^2 + (N_2-1)^2}{K\left(N_1+N_2-2\right)^2},$$

$$Q_N\left(\mathbf{U}_{K,N_1,N_2,J}, G_{K,N_1,N_2}\right) = 0.9504 > 0.9246 = 1 - \frac{1}{K\left(N_1+N_2-2\right)}J - \frac{2\left(N_1+N_2\right)^2}{\left(N_1+N_2-2\right)^2}J^{-1};$$

again (16) and (17) are satisfied.

4.2 Second Example

It may be argued that the results of Section 4.1 are only possible because we have used disconnected graphs. This is not the case. In this section we will illustrate the same deficiency using a family of connected graphs. We introduce the *connected* graphs H_{N_1,N_2} , H_{K,N_1,N_2} illustrated in Figure 2. Each H_{N_1,N_2} graph is a path of $N_1 + N_2$ nodes, with extra edges added between the first N_1 (resp. the second N_2) nodes at distance two of each other. The H_{K,N_1,N_2} is constructed by joining K H_{N_1,N_2} subgraphs in series.

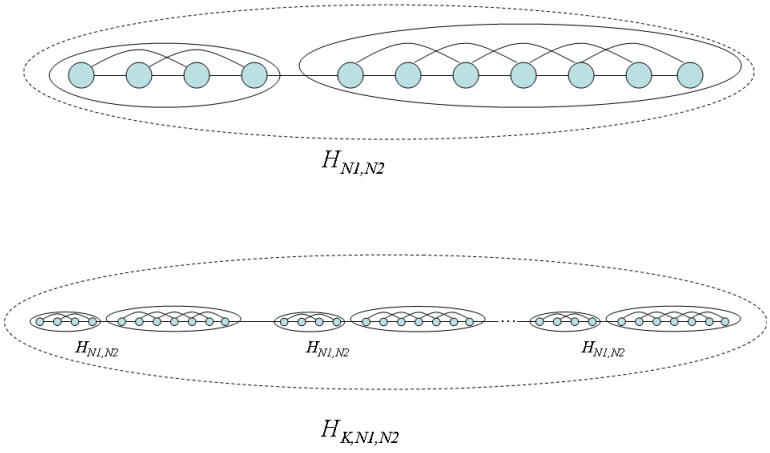

Figure 2: Graph Family B

We will use the same clusterings $\mathbf{V}_{K,N_1,N_2}$ and the sequences of clusterings $\{\mathbf{U}_{K,N_1,N_2J}\}_{J=1}^{\infty}$ as in Section 4.1. Once again, we claim that $\mathbf{V}_{K,N_1,N_2}$ is the natural clustering of H_{K,N_1,N_2}, for reasons similar to the ones discussed previously. Namely, cluster boundaries should occur across edges incident on the most weakly connected nodes; this shows that the $V_{K,N_1,N_2,k}$ clusters must be preserved; any partition of $V_{K,N_1,N_2,k}$ into finer clusters cannot be justified, since all of its nodes have the same connectivity pattern. Hence $\mathbf{V}_{K,N_1,N_2}$ is the "intuitively best" (i.e., the "natural") clustering of H_{K,N_1,N_2}.

Once again, we obtain a result similar to Theorem 4.3 in three steps.

Lemma 4.4 *For every $K, N_1, N_2 \in \mathbb{N}$ with $N_1, N_2 \geq 5$ we have*

$$Q_N\left(\mathbf{V}_{K,N_1,N_2}, H_{K,N_1,N_2}\right) < 1 - \frac{K\left(\left(4N_1-8\right)^2 + \left(4N_2-8\right)^2\right)}{\left(4K\left(N_1+N_2-2\right)\right)^2}. \tag{19}$$

Proof. We fix K, N_1, N_2 and, for brevity, we write H for H_{K,N_1,N_2} and $\mathbf{V}$ for $\mathbf{V}_{K,N_1,N_2}$; $\mathbf{V}'$ and $\mathbf{V}''$ have the same meaning as previously. For each $V_k \in \mathbf{V}'$, there are two border nodes on the left, two border nodes on the right and $N_1 - 4$ inner nodes. Each of the inner nodes has degree 4; each of the border nodes has degree 3, except for the first and last node of the graph, which have degree 2. Hence for each $V_k \in \mathbf{V}'$ we have the bounds

$$\left(N_1-4\right)\cdot 4 + 4\cdot 2 = 4N_1 - 8 < \deg\left(V_k\right) < 4N_1 - 4 = \left(N_1-4\right)\cdot 4 + 4\cdot 3.$$

Similarly, for each $V_k \in \mathbf{V}''$ we have the bounds

$$\left(N_2-4\right)\cdot 4 + 4\cdot 2 = 4N_2 - 8 < \deg\left(V_k\right) < 4N_2 - 4 = \left(N_2-4\right)\cdot 4 + 4\cdot 3.$$

Hence

$$K\left(\left(4N_1-8\right)^2 + \left(4N_2-8\right)^2\right) < \sum_{k=1}^{2K}\left(\deg\left(V_k\right)\right)^2 < K\left(\left(4N_1-4\right)^2 + \left(4N_2-4\right)^2\right). \tag{20}$$

The total number of edges is $m = \frac{\sum_{k=1}^{2K}\deg(V_k)}{2}$ and hence we have

$$\begin{gathered}\frac{K\left(4N_1-8+4N_2-8\right)}{2} < \frac{\sum_{k=1}^{2K}\deg\left(V_k\right)}{2} < \frac{K\left(4N_1-4+4N_2-4\right)}{2} \\ 2K\left(N_1+N_2-4\right) < m < 2K\left(N_1+N_2-2\right).\end{gathered} \tag{21}$$

In $\mathbf{V}_{K,N_1,N_2}$ there exist $2K-1$ extracluster edges, so we have

$$\frac{\sum_{k=1}^{2K}\left|E_k\right|}{m} < 1. \tag{22}$$

Combining (20) and (21) we get

$$\frac{\sum_{k=1}^{2K}\left(\deg\left(V_k\right)\right)^2}{\left(2m\right)^2} > \frac{K\left(\left(4N_1-8\right)^2 + \left(4N_2-8\right)^2\right)}{\left(4K\left(N_1+N_2-2\right)\right)^2}. \tag{23}$$

Finally, combining (21) and (23) we get the required bound. ■

Lemma 4.5 *For every $K, N_1, N_2, J \in \mathbb{N}$ with $N_1, N_2 \geq 3$ we have*

$$Q_N\left(\mathbf{U}_J, H_{K,N_1,N_2}\right) \geq 1 - \frac{3}{2K\left(N_1+N_2-4\right)}J - \frac{2\left(N_1+N_2\right)^2}{\left(N_1+N_2-4\right)^2}J^{-1}. \tag{24}$$

Proof. Each U_k has at most 6 extracluster edges, but these are counted twice; hence $\mathbf{U}_J$ cannot have more than $3J$ extra-cluster edges and so

$$\frac{\sum_{k=1}^{J+1}|E_k|}{m} \geq \frac{m-3J}{m} = 1 - \frac{3J}{m} > 1 - \frac{3J}{2K\left(N_1+N_2-4\right)}. \tag{25}$$

Each U_k has at most $\frac{N}{J} = \frac{K(N_1+N_2)}{J}$ nodes and each node has degree at most 4. Hence

$$\frac{\sum_{k=1}^{J+1}\left(\deg\left(U_k\right)\right)^2}{\left(2m\right)^2} \leq \frac{\left(J+1\right)\left(4\frac{K(N_1+N_2)}{J}\right)^2}{\left(4K\left(N_1+N_2-4\right)\right)^2} = \frac{2\left(N_1+N_2\right)^2}{\left(N_1+N_2-4\right)^2}J^{-1}. \tag{26}$$

Combining (25) and (26) we get the bound (24). ■

Hence, to ensure $Q_N\left(\mathbf{U}_{K,N_1N_2,J}, H_{K,N_1N_2}\right) > Q_N\left(\mathbf{V}_{K,N_1N_2}, H_{K,N_1N_2}\right)$ it suffices to choose appropriate K, N_1, N_2, J and use Lemmas 4.4 and 4.5. A sufficient condition, obtained from (19) and (24), is

$$1 - \frac{3}{2K\left(N_1+N_2-4\right)}J - \frac{2\left(N_1+N_2\right)^2}{\left(N_1+N_2-4\right)^2}J^{-1} > 1 - \frac{K\cdot\left(\left(4N_1-8\right)^2+\left(4N_2-8\right)^2\right)}{\left(4K\left(N_1+N_2-2\right)\right)^2}. \tag{27}$$

Now we can prove the following.

Theorem 4.6 *For every $K \in \mathbb{N}$ and $\varepsilon \in \left(0, \frac{1}{K}\right)$ there exist $N_1, N_2, J \in \mathbb{N}$ (depending on ε, K) such that*

$$Q_N\left(\mathbf{U}_{K,N_1,N_2,J}, H_{K,N_1,N_2}\right) > 1-\varepsilon > Q_N\left(\mathbf{V}_{K,N_1,N_2}, H_{K,N_1,N_2}\right) \tag{28}$$

$$\varepsilon > S\left(\mathbf{U}_{K,N_1,N_2,J}, \mathbf{V}_{K,N_1,N_2} | H_{K,N_1,N_2}\right). \tag{29}$$

Proof. Take any K and ε. Letting $N_1 = 6$, $J = xK$, $N_2 = x^2K$ we have

$$Q_N\left(\mathbf{U}_{K,N_1,N_2,J}, H_{K,N_1,N_2}\right) > 1 - \frac{3xK}{2K\left(2+x^2K\right)} - \frac{2\left(6+x^2K\right)^2}{xK\left(2+x^2K\right)^2},$$

$$Q_N\left(\mathbf{V}_{K,N_1,N_2}, H_{K,N_1,N_2}\right) < 1 - \frac{K\cdot\left(16^2+\left(4x^2K-8\right)^2\right)}{\left(4K\left(4+x^2K\right)\right)^2}.$$

For large enough x we have

$$1 - \frac{4+\left(x^2K-1\right)^2}{K\left(1+x^2K\right)^2} = 1 - \frac{1}{K} + o\left(x\right),$$

$$1 - \frac{x}{\left(1+x^2K\right)} - \frac{2\left(3+x^2K\right)^2}{\left(1+x^2K\right)^2xK} = 1 - \frac{3}{2Kx} - \frac{2}{Kx} + o\left(x\right) = 1 - \frac{7}{2Kx} + o\left(x\right).$$

Hence, for (28) to hold, x must be big enough for $o\left(x\right)$ to be negligible and we also need

$$1 - \frac{7}{2Kx} > 1 - \varepsilon > 1 - \frac{1}{K}$$

which is satisfied for any $x > \frac{7}{2K\varepsilon}$. It follows that (28) holds for every $K \in \mathbb{N}$ and $\varepsilon \in \left(0, \frac{1}{K}\right)$, if x is sufficiently big and $N_1 = 3$, $J = xK$, $N_2 = x^2K$. Hence we have proved (28). We can also prove that (29) is satified for big enough x; the argument is similar to the one used for (18) and hence is omitted. ■

Example. Here are two numerical examples.

1. Taking $K = 2$, $x = 5$ and $N_1 = 5$ we also get $J = xK = 10$ and $N_2 = x^2K = 50$. With these values we get

$$Q_N\left(\mathbf{V}_{K,N_1,N_2}, H_{K,N_1,N_2}\right) = 0.5538 < 0.5883 = 1 - \frac{K \cdot \left((4N_1 - 8)^2 + (4N_2 - 8)^2\right)}{\left(4K\left(N_1 + N_2 - 2\right)\right)^2},$$

$$Q_N\left(\mathbf{U}_{K,N_1,N_2,J}, H_{K,N_1,N_2}\right) = 0.7811 > 0.6203 = 1 - \frac{3J}{2K\left(N_1 + N_2 - 4\right)} - \frac{2\left(N_1 + N_2\right)^2}{\left(N_1 + N_2 - 4\right)^2} \cdot J^{-1};$$

   both (27) and (28) are satisfied.

2. Taking $K = 4$, $x = 10$ and $N_1 = 5$ we also get $J = xK = 40$ and $N_2 = x^2K = 400$. With these values we get

$$Q_N\left(\mathbf{V}_{K,N_1,N_2}, H_{K,N_1,N_2}\right) = 0.7527 < 0.7562 = 1 - \frac{K \cdot \left((4N_1 - 8)^2 + (4N_2 - 8)^2\right)}{\left(4K\left(N_1 + N_2 - 2\right)\right)^2},$$

$$Q_N\left(\mathbf{U}_{K,N_1,N_2,J}, H_{K,N_1,N_2}\right) = 0.9382 > 0.9116 = 1 - \frac{3J}{2K\left(N_1 + N_2 - 4\right)} - \frac{2\left(N_1 + N_2\right)^2}{\left(N_1 + N_2 - 4\right)^2} \cdot J^{-1};$$

   again both (27) and (28) are satisfied.

5 Discussion and Related Work

Combining Theorems 4.3 and 4.6 we see that for each of our G_{K,N_1,N_2} and H_{K,N_1,N_2} graphs there always exists a clustering $\mathbf{U}_{K,N_1,N_2,J}$ which achieves higher modularity than the natural clustering $\mathbf{V}_{K,N_1,N_2}$; also $\mathbf{U}_{K,N_1,N_2,J}$ is arbitrarily different from $\mathbf{V}_{K,N_1,N_2}$. It is conceivable that the maximum modularity is achieved at some other clustering $\mathbf{W}_{K,N_1,N_2}$ which is not very different from $\mathbf{V}_{K,N_1,N_2}$. However, until such a clustering is discovered, Theorems 4.3 and 4.6 indicate that, at least in some cases, the modularity Q_N is not a reliable clustering quality function.

In particular, recall that Q_N was introduced as "a measure for the strength of the community structure found by our algorithms, which gives us an objective metric for choosing the number of communities into which a network should be divided" [25]. However, we have examples of modularity maximization both underestimating (see [13, 14]) and overestimating (see Section 4) the true number of clusters.

We have just used the term "true number of clusters" and in earlier parts of this paper the terms "best clustering", "natural clustering", etc. However, as remarked by several researchers [12] these terms are, to a great extent, ambiguous. For "best" to have an objective meaning, it must be defined in terms of a cluster quality function; what is then a good quality function?

If our goal is to evaluate a particular quality function Q_1 (by checking that it selects the best clustering) then obviously we cannot use the same function to determine what is "best"; if we use another quality function Q_2 to determine "best", then we are only checking whether Q_1 and Q_2 agree. But this gets us no closer to determining whether they select the true clustering. The approach is rather self-referential.

Another possibility is to use labeled data, so that we know the true clustering in advance [36]. However, labeled data are not easy to get. In addition, there is always the possibility that labels (i.e., cluster membership) have been determined by external factors. Hoping that these factors will be sufficiently reflected in the connectivity of the graph seems overly optimistic. Indeed, for two of the most popular real world graph benchmarks (the Zachary karate club and football association network) it has been shown [35] that clusterings exist which have higher modularity than the true ones.

It appears that the only avenue left to evaluate clustering quality functions is to use graphs which have a "natural" (i.e., intuitively clear) community structure. This approach has been used in the construction of several benchmark graph familes [2, 20, 21] and is also the one we have used in Section 4.

The same problems arise in trying to estimate the "true number of clusters". At any rate, it is clear that in modularity maximization the selection of cluster number is effected (more or less accurately) by the Q_0 term. Several "modified modularity functions" have been proposed [3, 18, 22, 33, 26] which control the effect of the Q_0 term by multiplying it with a tuning parameter γ; but these variants may also be plagued by a resolution limit [5, 19, 34].

If we strip the Q_0 factor from Q_N we are left with the intracluster edge density Q_d which, by itself, does not provide a mechanism for cluster number selection. However this does not necessarily mean that Q_d cannot be used for community detection. Let us conclude with a discussion of possible enhancements which will enable the use of Q_d maximization for community detection.

The general approach we have in mind involves the maximization of Q_d for fixed values $K \in \{1, ..., K_{\max}\}$ and the subsequent selection of the optimal K by use of a postprocessing "*cluster number selection*" or "*model order*" or "*cluster validity*" criterion (these terms are almost, but not totally, synonymous). A plethora of such criteria has appeared in the "classical" pattern recognition literature. The interested reader can consult the classic book [11] or the review papers [10, 16, 17, 24]. It should be mentioned that in the pattern recognition community the cluster number selection problem has been recognized to be "a fundamental, and largely unsolved, problem in cluster analysis"[30].

Some of the "classical" criteria have also been used in the community detection literature, where alternatives to modularity maximization have been explored in conjunction with cluster number selection using criteria based on statistics and information theory. References to criteria of this type appear in [12], Section IX.B; for example the Akaike Information Criterion [1], the Bayesian Information Criterion [29], Minimum Description Length [27] etc. The gap statistic [32] and the knee criterion [28] should also be mentioned at this point.

Another way to estimate the number of clusters is through the use of some generative model, which essentially provides *prior knowledge* regarding the correct number of clusters. Some models of this type are discussed in [12], Section IX.A.

Yet another possibility is to use Q_d maximization in conjunction to some cluster validity criterion [6, 16, 17]. The idea is to perform Q_d maximization for $K \in \{1, 2, ..., K_{\max}\}$ and,

for each K value, grade the K-th clustering $\mathbf{V}^{(K)} = \left\{V_1^{(K)}, ..., V_K^{(K)}\right\}$ by a clustering quality function $Q(\mathbf{V}, G)$ of the form

$$Q\left(\mathbf{V}^{(K)}, G\right) = \left[\prod_{k=1}^{K} q\left(V_k^{(K)}, G\right)\right]^{1/K}, \tag{30}$$

where $q(V|G)$ is a cluster validity index (and the power $1/K$ is used to reduce the possibility that the $F(K) = \max_{\mathbf{V}\in\mathcal{V}_K} Q(\mathbf{V}, G)$ is decreasing with K). Hence (30) makes the "global" clustering quality $Q(\mathbf{V}, G)$ a "separable" function of the "local" validities of the clusters $V_1^{(K)}$, ..., $V_K^{(K)}$ and, hopefully, $F(K)$ attains a global maximum at $K = K_{true}$, the correct number of clusters. Similar approaches have been exploited in [15, 23, 31].

References

[1] H. Akaike, *A new look at the statistical model identification*, Automatic Control, IEEE Transactions on **19** (1974), no. 6, 716–723.

[2] R. Aldecoa and I. Marín, *Closed benchmarks for network community structure characterization*, Physical Review E **85** (2012), no. 2, 026109.

[3] J.I. Alvarez-Hamelin, B.M. Gastón, and J.R. Busch, *On weakly optimal partitions in modular networks*, Arxiv preprint arXiv:1008.3443 (2010).

[4] J.P. Bagrow, *Communities and bottlenecks: Trees and treelike networks have high modularity*, Physical Review E **85** (2012), no. 6, 066118.

[5] J.W. Berry, B. Hendrickson, R.A. LaViolette, and C.A. Phillips, *Tolerating the community detection resolution limit with edge weighting*, Physical Review E **83** (2011), no. 5, 056119.

[6] F. Boutin and M. Hascoet, *Cluster validity indices for graph partitioning*, Information Visualisation, 2004. IV 2004. Proceedings. Eighth International Conference on, IEEE, 2004, pp. 376–381.

[7] U. Brandes, D. Delling, M. Gaertler, R. Görke, M. Hoefer, Z. Nikoloski, and D. Wagner, *On finding graph clusterings with maximum modularity*, Graph-Theoretic Concepts in Computer Science, Springer, 2007, pp. 121–132.

[8] WYC Chen, AWM Dress, and WQ Yu, *Checking the reliability of a linear-programming based approach towards detecting community structures in networks*, IET systems biology **1** (2007), 286.

[9] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas, *Comparing community structure identification*, Journal of Statistical Mechanics: Theory and Experiment **2005** (2005), P09008.

[10] E. Dimitriadou, S. Dolničar, and A. Weingessel, *An examination of indexes for determining the number of clusters in binary data sets*, Psychometrika **67** (2002), no. 1, 137–159.

[11] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern classification and scene analysis 2nd ed.*, (1995).

[12] S. Fortunato, *Community detection in graphs*, Physics Reports **486** (2010), no. 3-5, 75–174.

[13] S. Fortunato and M. Barthelemy, *Resolution limit in community detection*, Proceedings of the National Academy of Sciences **104** (2007), no. 1, 36.

[14] B.H. Good, Y.A. de Montjoye, and A. Clauset, *Performance of modularity maximization in practical contexts*, Physical Review E **81** (2010), no. 4, 046106.

[15] R. Görke, A. Schumm, and D. Wagner, *Experiments on density-constrained graph clustering*, Arxiv preprint arXiv:1112.2143 (2011).

[16] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, Cluster validity methods: part I, ACM Sigmod Record **31** (2002), no. 2, 40–45.

[17] ———, Clustering validity checking methods: Part II, ACM Sigmod Record **31** (2002), no. 3, 19–27.

[18] G. Krings and V.D. Blondel, *An upper bound on community size in scalable community detection*, Arxiv preprint arXiv:1103.5569 (2011).

[19] J.M. Kumpula, J. Saramäki, K. Kaski, and J. Kertesz, *Limited resolution in complex network community detection with potts model approach*, The European Physical Journal B-Condensed Matter and Complex Systems **56** (2007), no. 1, 41–45.

[20] A. Lancichinetti, S. Fortunato, and F. Radicchi, *Benchmark graphs for testing community detection algorithms*, Physical Review E **78** (2008), no. 4, 046110.

[21] Z. Li, Y. Hu, B. Xu, Z. Di, and Y. Fan, *Detecting the optimal number of communities in complex networks*, Physica A: Statistical Mechanics and its Applications (2011).

[22] Z. Li, S. Zhang, R.S. Wang, X.S. Zhang, and L. Chen, *Quantitative function for community detection*, Physical review E **77** (2008), no. 3, 036109.

[23] J. Liu and T. Li, *A validity index approach for network partitions*, Physica A: Statistical Mechanics and its Applications (2011).

[24] G.W. Milligan and M.C. Cooper, *An examination of procedures for determining the number of clusters in a data set*, Psychometrika **50** (1985), no. 2, 159–179.

[25] M.E.J. Newman and M. Girvan, *Finding and evaluating community structure in networks*, Physical review E **69** (2004), no. 2, 026113.

[26] J. Reichardt and S. Bornholdt, *Statistical mechanics of community detection*, Physical Review E **74** (2006), no. 1, 016110.

[27] J. Rissanen, *Modeling by shortest data description*, Automatica **14** (1978), no. 5, 465–471.

[28] S. Salvador and P. Chan, *Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms*, Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on, IEEE, 2004, pp. 576–584.

[29] G. Schwarz, *Estimating the dimension of a model*, The annals of statistics **6** (1978), no. 2, 461–464.

[30] C.A. Sugar and G.M. James, *Finding the number of clusters in a dataset*, Journal of the American Statistical Association **98** (2003), no. 463, 750–763.

[31] G. Tibély, *Criterions for locally dense subgraphs*, Physica A: Statistical Mechanics and its Applications (2011).

[32] R. Tibshirani, G. Walther, and T. Hastie, *Estimating the number of clusters in a data set via the gap statistic*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) **63** (2001), no. 2, 411–423.

[33] VA Traag, P. Van Dooren, and Y. Nesterov, *Narrow scope for resolution-limit-free community detection*, Physical Review E **84** (2011), no. 1, 016114.

[34] J. Xiang and K. Hu, *Limitation of multi-resolution methods in community detection*, Arxiv preprint arXiv:1108.4244 (2011).

[35] B. Yang, D. Liu, and J. Liu, *Discovering communities from social networks: Methodologies and applications*, Handbook of Social Network Technologies and Applications (2010), 331–346.

[36] J. Yang and J. Leskovec, *Defining and evaluating network communities based on ground-truth*, Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics, ACM, 2012, p. 3.

[37] S. Zhang and H. Zhao, *Community identification in networks with unbalanced structure*, Physical Review E **85** (2012), no. 6, 066114.